Variability of Manual Segmentation of the Prostate in Axial T2-weighted MRI: A Multi-Reader Study

Becker, Anton S ; Chaitanya, Krishna ; Schawkat, Khoschy ; Mühlematter, Urs J ; Hötker, Andreas M ; Konukoglu, Ender ; Donati, Olivio F

Abstract: Purpose To evaluate the interreader variability in prostate and seminal vesicle (SV) segmentation on T2w MRI. Methods Six readers segmented the peripheral zone (PZ), transitional zone (TZ) and SV slice-wise on axial T2w prostate MRI examinations of n = 80 patients. Twenty different similarity scores, including dice score (DS), Hausdorff distance (HD) and volumetric similarity coefficient (VS), were computed with the VISCERAL EvaluateSegmentation software for all structures combined and separately for the whole gland (WG = PZ + TZ), TZ and SV. Differences between base, midgland and apex were evaluated with DS slice-wise. Descriptive statistics for similarity scores were computed. Wilcoxon testing to evaluate differences of DS, HD and VS was performed. Results Overall segmentation variability was good with a mean DS of 0.859 (±SD = 0.0542), HD of 36.6 (±34.9 voxels) and VS of 0.926 (±0.065). The WG showed a DS, HD and VS of 0.738 (±0.144), 36.2 (±35.6 vx) and 0.853 (±0.143), respectively. The TZ showed generally lower variability with a DS of 0.738 (±0.144), HD of 24.8 (±16 vx) and VS of 0.908 (±0.126). The lowest variability was found for the SV with DS of 0.884 (±0.0407), HD of 17 (±10.9 vx) and VS of 0.936 (±0.0509). We found a markedly lower DS of the segmentations in the apex (0.85 ± 0.12) compared to the base (0.87 ± 0.10, $p < 0.01$) and the midgland (0.89 ± 0.10, $p < 0.001$). Conclusions We report baseline values for interreader variability of prostate and SV segmentation on T2w MRI. Variability was highest in the apex, lower in the base, and lowest in the midgland.

DOI: https://doi.org/10.1016/j.ejrad.2019.108716

Journal Pre-proof

Variability of Manual Segmentation of the Prostate in Axial T2-weighted MRI: A Multi-Reader Study

Anton S. Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J. Mühlematter, Andreas M. Hötker, Ender Konukoglu, Olivio F. Donati

PII: S0720-048X(19)30366-3

DOI: https://doi.org/10.1016/j.ejrad.2019.108716

Reference: EURR 108716

To appear in: *European Journal of Radiology*

Received Date: 30 September 2019

Revised Date: 14 October 2019

Accepted Date: 16 October 2019

Please cite this article as: Becker AS, Chaitanya K, Schawkat K, Mühlematter UJ, Hötker AM, Konukoglu E, Donati OF, Variability of Manual Segmentation of the Prostate in Axial T2-weighted MRI: A Multi-Reader Study, *European Journal of Radiology* (2019), doi: https://doi.org/10.1016/j.ejrad.2019.108716

Variability of Manual Segmentation of the Prostate in Axial T2-weighted MRI: A Multi-Reader Study

Original Research

EJR - Original Article

Running title: Interreader Variability in Prostate Segmentation

Anton S. Becker, M.D., Ph.D.*[1,2]; Krishna Chaitanya, M.Sc.[3]; Khoschy Schawkat, M.D.[1,4]; Urs J. Mühlematter, M.D.[1]; Andreas M. Hötker, M.D.[1]; Ender Konukoglu, Ph.D.[3]; Olivio F. Donati, M.D.[1]

[1] Institute of Diagnostic and Interventional Radiology, University Hospital Zurich; Zurich, Switzerland

[2] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York City, USA

[3] Computer Vision Laboratory, Department of Information Technology and Electrical Engineering, ETH Zurich

[4] Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, USA

*Correspondence to: Anton S. Becker, Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, 8091 Zürich, Switzerland. E-Mail: anton.becker@usz.ch

Highlights

- Prostate MRIs from different institutions of 80 patients were manually, voxel-wise segmented.
- Six readers (2 trainees, 2 radiologists, 2 researchers) performed the segmentations.
- Variability was highest in the apex, lower in the base, and lowest in the midgland.

Abstract

Purpose: To evaluate the interreader variability in prostate and seminal vesicle (SV) segmentation on T2w MRI.

Methods: Six readers segmented the peripheral zone (PZ), transitional zone (TZ) and SV slice-wise on axial T2w prostate MRI examinations of n=80 patients.

Twenty different similarity scores, including dice score (DS), Hausdorff distance (HD) and volumetric similarity coefficient (VS), were computed with the VISCERAL EvaluateSegmentation software for all structures combined and separately for the whole gland (WG=PZ+TZ), TZ and SV. Differences between base, midgland and apex were evaluated with DS slice-wise.

Descriptive statistics for similarity scores were computed. Wilcoxon testing to evaluate differences of DS, HD and VS was performed.

Results: Overall segmentation variability was good with a mean DS of 0.859 (±SD = 0.0542), HD of 36.6 (±34.9 voxels) and VS of 0.926 (±0.065). The WG showed a DS, HD and VS of 0.738 (±0.144), 36.2 (±35.6 vx) and 0.853 (±0.143), respectively. The TZ showed generally lower variability with a DS of 0.738 (±0.144), HD of 24.8 (±16 vx) and VS of 0.908 (±0.126). The lowest variability was found for the SV with DS of 0.884 (±0.0407), HD of 17 (±10.9 vx) and VS of 0.936 (±0.0509). We found a markedly lower DS of the segmentations in the apex (0.85±0.12) compared to the base (0.87±0.10, $p<0.01$) and the midgland (0.89±0.10, $p<0.001$).

Conclusions: We report baseline values for interreader variability of prostate and SV segmentation on T2w MRI. Variability was highest in the apex, lower in the base, and lowest in the midgland.

Keywords: segmentation, prostate, zonal anatomy, interreader agreement, variability

Introduction

Accurate prostate segmentation is crucial in the planning of biopsies (i.e. MR-ultrasound image fusion), focal ablative treatments of localized prostate cancer (1) or irradiation of inoperable tumors (2). Moreover, the definition of prostate boundaries on MR images are a crucial step for the radiologist to detect extracapsular extension or invasion of adjacent structures, such as the neurovascular bundle, rectum or seminal vesicles (3).

The segmentation of the T2-weighted (T2w) MRI images is usually performed in a manual, slice-by-slice fashion. Because this is a tedious and time-consuming process, recent research efforts have focused on automating prostate segmentation (4), e.g. using deep learning algorithms (5). Since the inception of the PROMISE challenge in 2012 (5), there has been a perpetual chase for higher similarity scores to the “ground truth segmentation”. However, while novel deep-learning architectures promise to deliver incremental improvements (6), it is not clear whether these simply reflect the algorithm learning the particular reference segmentation of a single reader at a single point in time (7), or whether they reflect a truly meaningful improvement towards the human baseline. In fact, this very baseline is not yet known: Only very few studies have systematically investigated the underlying human interreader variability in prostate segmentation (8, 9), and no study has been undertaken to evaluate the delineation of the different anatomical zones of the prostate and the seminal vesicles. Both are crucial steps on the path to

automated prostate cancer detection and local staging algorithms (10). Furthermore, it is unclear how different levels of expertise/experience influence segmentation accuracy or interreader variability.

Hence, the purpose of this study was to determine the interreader variability of radiologists with different levels of experience and non-radiologists in the segmentation of the prostate in T2w MR images, and to provide a publicly accessible dataset of MRI images and six candidate segmentations.

Methods

Datasets

We compiled a cohort of 80 patients from two different cohorts/datasets. Dataset I (in-house) consisted of 68 treatment-naïve patients from a previously published retrospective cohort study (11). Institutional review board approval for the use of these data had been obtained under a waiver for additional informed consent. Dataset II consisted of 12 patients from the publicly available “Prostate-3T” dataset from the cancer imaging archive (TCIA, direct link: http://bit.ly/prostate3t). The following 12 IDs were randomly selected: 09, 13, 15, 19, 22, 31, 33, 37, 38, 40, 49, 53. The corresponding segmentations can be downloaded at BLINDED FOR REVIEW. Image quality of both cohorts was comparable.

Segmentation

For segmentation we used the freely available ITK-Snap software (v. 3.6.0; itksnap.org) in *polygon* mode with the standard settings (segment length 8 px). Six readers

performed the segmentations: two experienced readers, Exp1 (BLINDED) and Exp2 (BLINDED) (board-certified radiologists with 6 and 5 years of experience, respectively), two radiology residents in training: Res1 (BLINDED) and Res2 (BLINDED) (5^{th} and 3^{rd} year of the 5-year residency), two computer vision scientists, Sci1 (BLINDED) and Sci2 (BLINDED) (non-radiologists; associate professor and PhD student) with over 14 and 2 years of experience in medical image segmentation. The basic prostatic anatomy is reviewed in **Figure 1 a**/**b**. Readers were instructed to first segment the whole peripheral zone (PZ) with the edges well overlapping with the transitional zone and seminal vesicles in order to avoid pixels with the *background* label on the edges of the segmentation. The readers then segmented the transitional zone (TZ) and seminal vesicles (SV), drawing over the PZ label where needed. This process is depicted in **Figure 2**. Since many patients at risk for PCa are older and suffer concurrent benign prostatic hyperplasia, their anatomy is usually to a greater or lesser degree distorted as shown in **Figure 1 c-f**. This led us to not segment the central zone (CZ) separately, since it is usually compressed, displaced and not well delineated. Furthermore, only around 10% of cancers arise from this area. Henceforth, we implicitly include the CZ whenever we refer to the TZ unless otherwise stated. Furthermore, the anterior fibromuscular stroma was not segmented separately. It is usually compressed and deformed to a sliver, and its borders are ill-defined. Also, cancer does usually not arise from this structure due to its lack of glandular elements.

Similarity metrics and statistical analysis

The open-source *VISCERAL EvaluateSegmentation* software (Apache License v2), available for download at https://github.com/Visceral-Project/EvaluateSegmentation was used for computation of the similarity of the segmentations. The software implements

many different similarity metrics as published in detail in (12). Since each metric is sensitive to a different quality aspect of 3D segmentation, we report 16 different metrics, calculated pairwise and summarized as mean and standard deviation. However, since many of those parameters exhibit strong co-correlations, statistical testing is only performed on the three most commonly reported metrics in the literature (Dice-score (DS), Hausdorff distance (HD) and the volumetric similarity coefficient (VS)) in order to avoid Type I errors. To evaluate the differences between base, midgland and apex of the prostate (i.e. divided by thirds in craniocaudal axis), the dice score was calculated for every slice and subsequently summarized for each third in every prostate/reader. A paired wilcoxon-test was used for comparisons and the p-values were corrected with the Bonferroni-Holm procedure for the number of reader pairs/combinations. A p-value <0.05 was considered indicative of a statistically significant difference.

Results

Overall segmentation variability was good with a mean DS of 0.859 (±SD = 0.0542), HD of 36.6 (±34.9 px) and VSC of 0.859 (±0.0542). When evaluating only the gland itself (TZ+PZ) without the SV, we found that the DS 0.738 (±0.144) and VS 0.853 (±0.143) markedly decreased, while the HD remained virtually the same 36.2 (±35.6 px). The TZ showed generally lower variability when separated from the PZ with a DS of 0.738 (±0.144), HD of 24.8 (±16 px) and VS of 0.908 (±0.126). The lowest variability was found for segmentation of the SV with high DS 0.884 (±0.0407), low HD 17 (±10.9 px) and high VS 0.936 (±0.0509). These results as well as the remaining metrics are summarized in **Table 1**. There were several significant differences between radiologists and non-

radiologists as summarized in **Table 2**: Only the radiology residents (representing intermediate experience) exhibited one non-significant difference with each of the other category (representing low and high experience, respectively). The full pair-wise data can be found online in the **supplementary Tables**.

To further investigate the high variability in the whole gland/PZ, we computed the DS for every slice of the segmentation. To account for the different number of slices in the exams, we summarized the DS for every third of the prostate (in the z-axis): Apex, midgland and base We found a markedly lower similarity of the segmentations in the apex (0.85±0.12) compared to the base (0.87±0.10, $p<0.01$) and the midgland (0.89±0.10, $p<0.001$) as can be appreciated in **Figure 3**. Note that the 3D-DS and 2D-DS are not directly comparable.

Discussion

We systematically analyzed the interreader variability of human readers in zonal segmentation of the prostate gland and seminal vesicles based on T2w MRI. We found low variability for the SV, slightly higher variability for the TZ and significant higher variability in the apical and basal portions of the prostate gland/PZ. The latter results are partly in line with a previously published, smaller study by Shahedi et al. (9) who reported a lower Dice-score in the base and apex (0.67/0.66). In line with our results, they also reported a lower variability in the midgland (0.88), however, they did not find a difference between apex and base, which may be an artifact of the relatively small number of subjects (n=10). On the other hand, Wang et al. (8) studied the segmentation of the whole prostate in 3 readers and n=40 subjects and found a comparable DS (0.84)

to our 2D values, however, they did not further analyse in which regions/height the variability was higher. Furthermore, an earlier study by Martin et al. (4) found a nearly identical variability, which was improved by helping the reader with a semi-automatic algorithm. Lastly, we found significant differences between radiologists and non-radiologists, which suggests that, in contrast to the findings of (13), an interreader-baseline of non-radiologists may not suffice for meaningful comparison to new segmentation algorithms.

With increasing performance of various machine learning algorithms (in particular deep neural networks), there is great hope for algorithms which help the radiologist detect prostate cancer in MRI (14). This is of interest not only for large centers, where such algorithms may enable higher examination throughput, but also for smaller facilities where the number of prostate MRI examinations is low and thus a broader range of pathologies needs to be covered overall. However, it is important that such algorithms, implicitly or explicitly, have an understanding of the zonal anatomy of the prostate. This is important for at least two reasons: First, the cancer prevalence is very different amongst the anatomical zones, with the PZ harboring the majority (~70-80%) of cancers. In Bayesian terms, this results in a different prior probability for candidate lesions in the different zones. Moreover, TZ and PZ cancers exhibit different imaging characteristics and should thus be differently evaluated by the reading radiologist (15). A recent study by Antonelli et al. suggests that the same holds true for machine learning algorithms, which exhibit different performance characteristics depending on the zonal origin of the cancer (16). Second, local staging is an important part of reading a prostate MRI. In order to evaluate extracapsular extension or invasion of neighboring structures, e.g. the SV, the boundaries of these structures need to be precisely defined. While it is

often said that in theory, with unlimited, well-labelled data, an optimal algorithm should be able to automatically develop internal representations of these concepts. However, in the real world good data is sparse and it makes sense to incorporate prior knowledge: For example, shape priors have been shown to assist the deep learning based segmentation algorithms in other body regions such as the kidney or the heart (17, 18).

The reason for the relative sparsity of high-quality labelled training data is the same as why automatic segmentation algorithms for the prostate are important: Manual segmentation is extremely time-consuming. On the other hand, segmentation of the prostate is also an extremely important task and has to be performed frequently in the clinical routine, i.e. for MRI-ultrasound image fusion in order to obtain targeted biopsy samples. It is not uncommon for patients to undergo multiple biopsies before the diagnosis of a clinically significant cancer can be established (19). Each time, the prostate needs to be newly segmented. Furthermore, when receiving external beam radiation therapy (EBRT), the treatment is usually delivered in daily fractions over a period of multiple weeks, requiring daily new segmentations (20). In the case of EBRT reliable and accurate segmentation is of utmost importance in order to avoid radiation toxicity to surrounding structures. Hence, reliable automatic prostate segmentation algorithms may not only save time but may also help improve patient oucomes.

Our study has several limitations that need to be mentioned. Although we provided a comprehensive evaluation of the segmentation of the prostate with all established similarity metrics, we did not evaluate the accuracy in segmenting other clinically important structures, such as the urethra or the neurovascular bundles. Hence, we cannot conclusively assess whether this would have been accurately assessed e.g. by

the non-radiologists. However, some of these structures are often very hard to delineate even for an experienced radiologist, and may only be identifiable on a given slice with the information from the slice above and/or below. It stands to reason that unless better MRI sequences become available which allow more confident delineation, automatic algorithms will not be able to reliably segment these structures. Another limitation is that we are not able to share our full dataset due to patient confidentiality/data protection. However, we attempted to overcome this limitation by including a second cohort from a publicly available dataset, the segmentations of which are publicly available for further research. Furthermore, the difference in agreement between base and apex was, although statistically significant, fairly small. Further studies are necessary to determine whether this difference is clinically relevant. Lastly, although some of our patients exhibited MR visible cancer lesions, we did not further evaluate the influence on the segmentation. The rationale was that computer algorithms will also be evaluated on a mixed, real-world cohort; hence our sample is more representative of the clinical routine. As for the segmentation of cancer lesions, we believe this is a separate matter that deserves a dedicated study.

In summary, we report baseline values for interreader variability of prostate segmentation in MRI. Variability was highest in the apex, lower in the base, and lowest in the midgland.

Grant support: ASB was supported by the Prof. Dr. Max Cloëtta Foundation, the medAlumni UZH and the Swiss Society of Radiology.

Conflicts of Interest

The authors of this manuscript declare no relevant conflicts of interest, and no relationships with any companies, whose products or services may be related to the subject matter of the article.

Acknowledgements and Conflicts of Interest: None

References

1. Garnier C, Bellanger J-J, Wu K, et al.: Prostate segmentation in hifu therapy. *IEEE Transactions on Medical Imaging* 2011; 30:792–803.

2. Guckenberger M, Meyer J, Baier K, Vordermark D, Flentje M: Distinct effects of rectum delineation methods in 3D-confromal vs. imrt treatment planning of prostate cancer. *Radiation Oncology* 2006; 1:34.

3. Vargas SO, Jiroutek M, Welch WR, Nucci MR, D'Amico AV, Renshaw AA: Perineural invasion in prostate needle biopsy specimens: Correlation with extraprostatic extension at resection. *American journal of clinical pathology* 1999; 111:223–228.

4. Martin S, Rodrigues G, Patil N, et al.: A multiphase validation of atlas-based automatic and semiautomatic segmentation strategies for prostate mri. *International Journal of Radiation Oncology* Biology* Physics* 2013; 85:95–100.

5. Litjens G, Toth R, Ven W van de, et al.: Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Medical image analysis* 2014; 18:359–373.

6. Zhu Y, Wei R, Gao G, et al.: Fully automatic segmentation on prostate mr images based on cascaded fully convolution network. *Journal of Magnetic Resonance Imaging* 2019; 49:1149–1156.

7. Chaitanya K, Karani N, Baumgartner CF, Becker A, Donati O, Konukoglu E: Semi-supervised and task-driven data augmentation. *Information Processing in Medical Imaging* 2019:29–41.

8. Wang B, Lei Y, Tian S, et al.: Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic mri prostate segmentation. *Medical physics* 2019.

9. Shahedi M, Cool DW, Bauman GS, Bastian-Jordan M, Fenster A, Ward AD: Accuracy validation of an automated method for prostate segmentation in magnetic resonance imaging. *Journal of digital imaging* 2017; 30:782–795.

10. Muehlematter UJ, Burger IA, Becker AS, Schawkat K, Hötker AM, Reiner CS, Müller J, Rupp NJ, Rüschoff JH, Eberli D, Donati OF: Diagnostic Accuracy of Multiparametric MRI versus 68Ga-PSMA-11 PET/MRI for Extracapsular Extension and Seminal Vesicle Invasion in Patients with Prostate Cancer. *Radiology*. 2019 Sep 10:190687.

11. BLINDED

12. Taha AA, Hanbury A: Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC medical imaging* 2015; 15:29.

13. Bezinque A, Moriarity A, Farrell C, Peabody H, Noyes SL, Lane BR: Determination of prostate volume: A comparison of contemporary methods. *Academic radiology* 2018; 25:1582–1587.

14. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, Kuder TA, Stenzinger A, Hohenfellner M, Schlemmer HP, Maier-Hein KH: Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology*. 2019 Oct 8:190938.

15. Turkbey B, Rosenkrantz AB, Haider MA, et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology* 2019.

16. Antonelli M, Johnston EW, Dikaios N, et al.: Machine learning classifiers can predict gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *European Radiology* 2019.

17. Ravishankar H, Venkataramani R, Thiruvenkadam S, Sudhakar P, Vaidya V: Learning and incorporating shape models for semantic segmentation. In *International conference on medical image computing and computer-assisted intervention*. Springer; 2017:203–211.

18. Oktay O, Ferrante E, Kamnitsas K, et al.: Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging* 2018; 37:384–395.

19. Hambrock T, Somford DM, Hoeks C, et al.: Magnetic resonance imaging guided prostate biopsy in men with repeat negative biopsies and increased prostate specific antigen. *The Journal of urology* 2010; 183:520–528.

20. Hegemann N-S, Guckenberger M, Belka C, Ganswindt U, Manapov F, Li M: Hypofractionated radiotherapy for prostate cancer. *Radiation oncology* 2014; 9:275.

21. McNeal JE: The zonal anatomy of the prostate. *The prostate* 1981, 2(1), 35-49.

Figure Legends

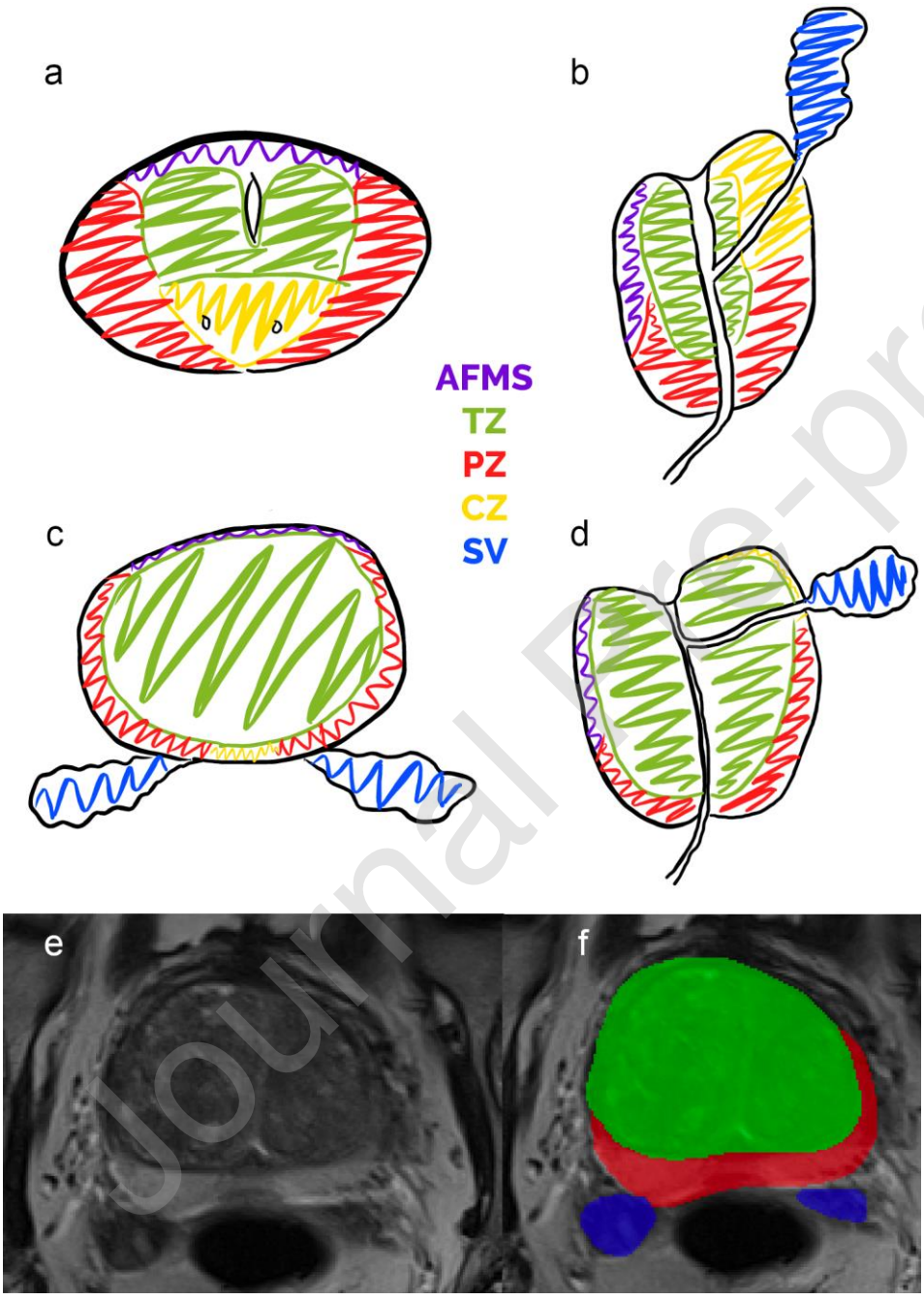

Figure 1: Zonal anatomy of the prostate as established by J.E. McNeal (21) in an axial slice through the base **(a)** and a midsagittal section **(b)**. More often than not, this

anatomy is heavily distorted by benign prostatic hyperplasia (of the TZ), which compresses and displaces CZ and AFMS to small, ill-defined slivers of tissue **(c&d)**. Hence, only PZ, TZ and SV were labelled in the segmentation process **(e&f)**. Abbr.: TZ=transition zone, PZ=peripheral zone, CZ=central zone, AFMS=anterior fibromuscular stroma, SV=seminal vesicles.

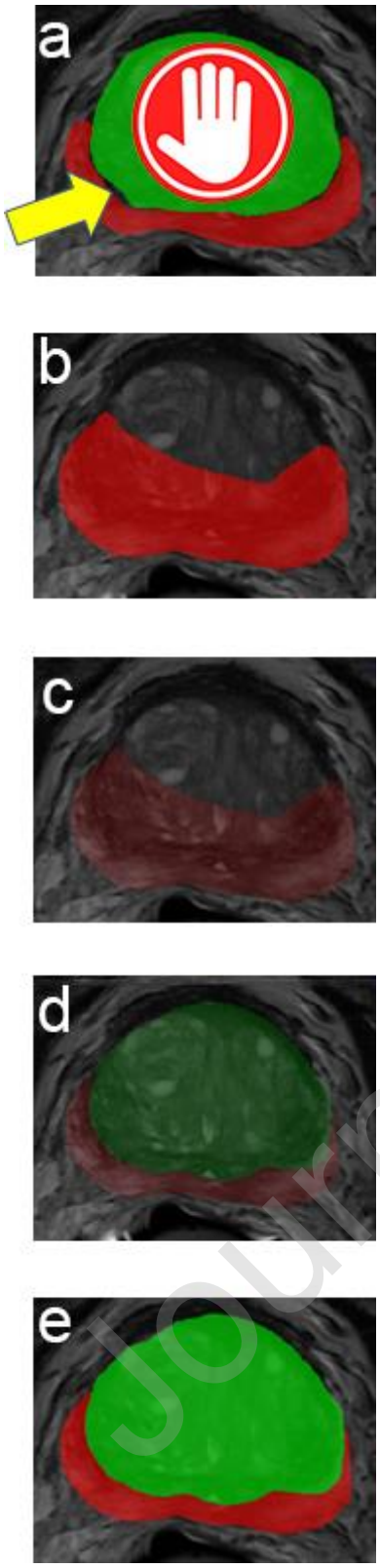

Figure 2: To avoid background labels within the prostate (yellow arrow in (a)), the readers were instructed to follow the sequence depicted in (b)-(d) i.e. to first segment

the PZ labelling a large part of the TZ as well, and subsequently re-label these falsely assigned voxels with the true TZ label (e) for a gapless segmentation.

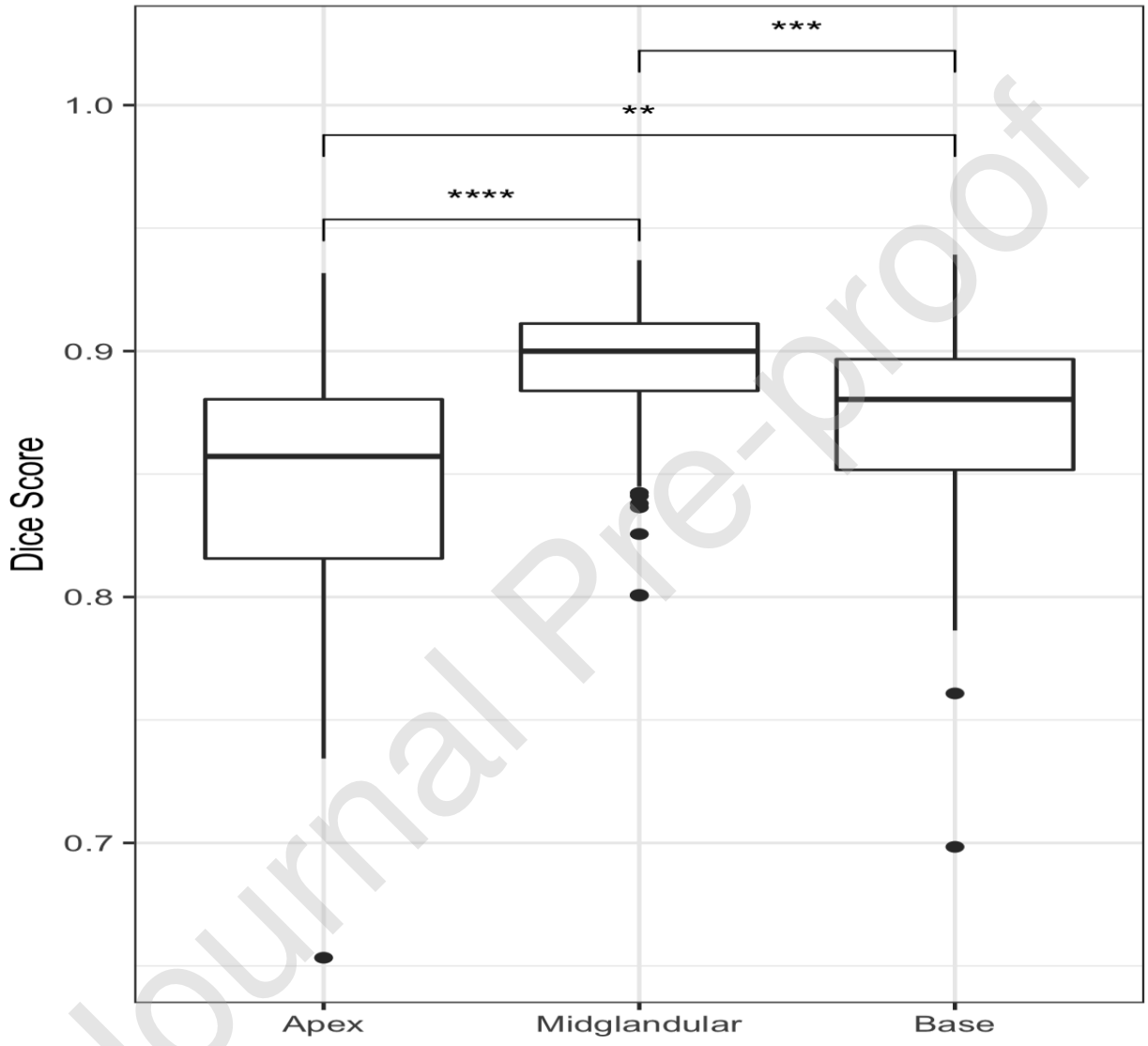

Figure 3: Slice-wise Dice score shows siginficantly higher variability in the apical third (0.85±0.12) compared to the base (0.87±0.10, **p<0.01). Both base and apex exhibited higher variability compared to the midgland (0.89±0.10, ***p<0.001 and ****p<0.0001).

Tables

Table 1: Summarized pairwise similarity metrics for all readers and radiologists only for all structures (All), seminal vesicles (SV), transition zone (TZ) and TZ+peripheral zone (Whole gland). Radiologists refers to the four radiologists, two of which were board-certified and two in training.

Pair	Structure	Accuracy	Adjusted Rand Index	Area under ROC Curve	Average Hausdorff Distance (in voxel)	Hausdorff Distance (in voxel)	Cohen Kappa	Dice Coefficient (F1-Measure)	F-Measure	Interclass Correlation	Jaccard Coefficient
All Readers	All	0.0077 (±0.00604)	0.842 (±0.0563)	0.946 (±0.0328)	0.498 (±1.42)	36.6 (±34.9)	0.853 (±0.0551)	0.859 (±0.0542)	0.988 (±0.00579)	0.859 (±0.0542)	0.756 (±0.0776)
Radiologists	All	0.0063 (±0.004)	0.855 (±0.0392)	0.941 (±0.0311)	0.317 (±0.395)	33.1 (±23.6)	0.866 (±0.0391)	0.871 (±0.039)	0.989 (±0.00433)	0.871 (±0.039)	0.774 (±0.0587)
All Readers	SV	0.00501 (±0.00409)	0.873 (±0.0417)	0.956 (±0.0289)	0.201 (±0.121)	17 (±10.9)	0.88 (±0.0411)	0.884 (±0.0407)	0.992 (±0.00386)	0.884 (±0.0407)	0.795 (±0.0631)
Radiologists	SV	0.00415 (±0.00315)	0.881 (±0.0314)	0.951 (±0.0284)	0.192 (±0.117)	16.9 (±8.56)	0.888 (±0.0312)	0.892 (±0.0311)	0.993 (±0.00346)	0.892 (±0.0311)	0.806 (±0.0496)
All Readers	TZ	0.00384 (±0.00406)	0.787 (±0.204)	0.909 (±0.106)	0.851 (±1.87)	24.8 (±16)	0.792 (±0.205)	0.796 (±0.204)	0.993 (±0.00687)	0.796 (±0.204)	0.694 (±0.2)
Radiologists	TZ	0.00381 (±0.00502)	0.766 (±0.239)	0.89 (±0.121)	1.07 (±2.25)	27.6 (±17)	0.771 (±0.24)	0.774 (±0.239)	0.992 (±0.00812)	0.774 (±0.239)	0.675 (±0.228)
All Readers	Whole gland	0.00252 (±0.0027)	0.733 (±0.144)	0.905 (±0.0757)	1.57 (±5.28)	36.2 (±35.6)	0.736 (±0.144)	0.738 (±0.144)	0.996 (±0.00296)	0.738 (±0.144)	0.602 (±0.153)
Radiologists	Whole gland	0.00219 (±0.00166)	0.748 (±0.138)	0.897 (±0.076)	1.15 (±3.02)	33.9 (±24.3)	0.751 (±0.138)	0.752 (±0.139)	0.996 (±0.00201)	0.752 (±0.139)	0.619 (±0.145)

Pair	Structure	Accuracy	Adjusted Rand Index	Area under ROC Curve	Average Hausdorff Distance (in voxel)	Hausdorff Distance (in voxel)	Cohen Kappa	Dice Coefficient (F1-Measure)	F-Measure	Interclass Correlation	Jaccard Coefficient

Pair	Structure	Mahanabolis Distance	Mutual Information	Global Consistency Error	Precision (Confidence)	Probabilistic Distance	Variation of Information	Volumetric Similarity Coefficient	Rand Index
All Readers	All	0.158 (±0.12)	0.182 (±0.0707)	0.0211 (±0.00956)	0.859 (±0.0542)	0.000667 (±0.000343)	0.142 (±0.0539)	0.926 (±0.0652)	0.977 (±0.0111)
Radiologists	All	0.131 (±0.0822)	0.19 (±0.0709)	0.0203 (±0.00805)	0.871 (±0.039)	0.000589 (±0.00022)	0.139 (±0.0484)	0.942 (±0.0428)	0.979 (±0.00845)
All Readers	SV	0.145 (±0.0865)	0.164 (±0.0706)	0.0142 (±0.0069)	0.884 (±0.0407)	0.000523 (±0.000217)	0.1 (±0.0423)	0.936 (±0.0509)	0.985 (±0.00757)
Radiologists	SV	0.13 (±0.0717)	0.168 (±0.0708)	0.0138 (±0.00642)	0.892 (±0.0311)	0.000481 (±0.000158)	0.0985 (±0.0407)	0.949 (±0.0367)	0.985 (±0.0068)
All Readers	TZ	0.309 (±0.39)	0.106 (±0.0743)	0.012 (±0.00883)	0.796 (±0.204)	0.0114 (±0.0741)	0.0845 (±0.0507)	0.908 (±0.126)	0.986 (±0.0132)
Radiologists	TZ	0.336 (±0.465)	0.105 (±0.0763)	0.0129 (±0.00999)	0.774 (±0.239)	0.0185 (±0.101)	0.09 (±0.0562)	0.903 (±0.146)	0.985 (±0.0155)
All Readers	Whole gland	0.36 (±0.43)	0.0441 (±0.0259)	0.00677 (±0.00396)	0.738 (±0.144)	0.0945 (±2.25)	0.051 (±0.0248)	0.853 (±0.143)	0.992 (±0.00574)
Radiologist	Whole	0.348	0.0469	0.00669	0.752	0.0634	0.0505	0.877	0.992 (±0.00398

Pair	Structure	Mahanabolis Distance	Mutual Information	Global Consistency Error	Precision (Confidence)	Probabilistic Distance	Variation of Information	Volumetric Similarity Coefficient	Rand Index
s	gland	(±0.496)	(±0.03)	(±0.00325)	(±0.139)	(±1.15)	(±0.0215)	(±0.118)	)

Table 2: Adjusted *p*-values of the similarity metrics between different reader parings, showing significant differences for most combinations. Radiology residents (representing intermediate experience) exhibited one non-significant difference with each of the other category (representing low and high experience, respectively).

Comparison	Dice Coefficient (F1-Measure)	Volumetric Similarity Coefficient	Hausdorff Distance (in voxel)
Non vs. Expert Radiologists	< 0.001	< 0.001	0.040
Non vs. Resident Radiologists	0.37	< 0.001	< 0.001
Resident vs. Expert Radiologists	< 0.001	0.25	< 0.001
Non-Rad.-Radiologists vs. Radiologists only	0.002	< 0.001	0.028